ICS 35.240.99
CCS L70

T/SAIAS

上海市人工智能行业协会团体标准

T/SAIAS XXX—XXXX

文博语料库建设导则

Guidelines for the Construction of Cultural and Museum Corpora

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

上海市人工智能行业协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由上海市人工智能行业协会提出并归口。

本文件起草单位：

本文件主要起草人：

本标准首次制定。

首期执行单位：

引　言

人工智能是新一轮科技革命和产业变革的重要驱动力量，高质量、大规模、多模态的语料库则是人工智能技术研发与创新应用不可或缺的基础性战略资源。在文博领域，构建符合文化遗产保护、研究、展示与传播特殊需求的专业语料库，对于推动人工智能技术在该领域的深度融合、构建智慧文博生态体系、培育文化领域新质生产力具有至关重要的意义。制定《人工智能文博语料库建设导则》，旨在系统引导和规范文博语料资源的建设、治理与应用全流程，为人工智能赋能文化遗产事业提供标准化支撑。

国家层面，《关于加强文物科技创新的意见》、《“十四五”文物保护和科技创新规划》等一系列政策文件，均强调运用人工智能等新一代信息技术加强文物数字化保护与活化利用，并明确提出构建国家文物资源大数据库、推动文物数据开放共享等任务，为文博语料库的建设指明了方向。《“数据要素X”三年行动计划（2024-2026年）》等政策进一步鼓励文化数据资源开发与利用，支持文化大模型发展，为文博语料库的应用价值释放提供了广阔空间。

上海市层面，积极落实国家战略，先后出台了《上海市打造文旅元宇宙新赛道行动方案（2023-2025年）》、《上海市推进“人工智能+”高质量发展行动方案（2025）》等政策，明确要求挖掘文旅元宇宙应用场景，推动人工智能大模型技术与文化产业深度融合，支持建设高质量行业语料库。同时，相关地方法规也强调运用数字化、智能化手段加强文物保护与传承，为文博语料库的建设提供了坚实的制度保障。上海文博机构在藏品数字化、智慧博物馆建设方面已积累了丰富经验，为语料库建设奠定了实践基础。

本导则积极响应国家文化数字化战略和上海市人工智能产业发展要求，充分借鉴现有实践成果，聚焦文博领域的知识特殊性与数据多样性，系统规范了文博语料的多模态构成、核心知识体系、采集清洗、语义标注、安全伦理与治理机制等关键环节的技术要求与操作规范。我们期待本导则的实施能够有效提升文博语料库建设的规范化、专业化水平，促进文博数据的开放共享、语义互联与高效利用，助力构建协同发展的智慧文博生态，为传承中华优秀传统文化、提升文化自信注入新动能。

文博语料库建设导则

1 范围

本文件规定了文物与博物馆领域语料库建设的一般要求、核心流程和技术方法，涵盖文博语料的定义、采集、清洗、标注、管理、应用及安全等全生命周期环节。

本文件适用于博物馆、考古研究机构、文物保护管理机构及企业等单位，针对可移动文物与不可移动文物，开展文物与博物馆领域多模态语料库的规划、设计、开发、维护及评估等相关工作。其他文化遗产相关领域的语料库建设亦可参照本文件执行。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 4894-2009 信息与文献 术语

GB/T 37965-2019 信息与文献 文化遗产信息交换的参考本体

GB/T 42755-2023 人工智能 面向机器学习的数据标注规程

GB/T 43697—2024 数据安全技术 数据分类分级规则

GB/T 44206-2024 馆藏文物病害数据库建设规范

T/SAIAS 015—2024 语料库建设导则

3 术语和定义

T/SAIAS 015—2024界定的以及下列术语和定义适用于本文件。

3.1

文博语料 cultural and museum corpus

与可移动文物（如器物、书画、典籍等）、不可移动文物（如古遗址、古墓葬、古建筑、石窟寺、石刻、近现代重要史迹及代表性建筑等）、博物馆、非物质文化遗产相关的，用于计算和分析的多模态语言材料样本，包括但不限于文本、图像、音频、视频、三维模型及其相关元数据。

3.2

文博语料库 cultural and museum corpora

依据特定目的和方法，系统采集、加工、标注和组织的文博语料构成的电子数据库。

4 缩略语

下列缩略语适用于本文件。

ASR 自动语音识别（Automatic speech recognition）

CIDOC CRM 文化遗产概念参考模型（CIDOC Conceptual Reference Model）

GDPR 通用数据保护条例（General Data Protection Regulation）

OCR 光学字符识别（Optical Character Recognition）

URI 统一资源标识符（Uniform Resource Identifier）

5 文博语料

5.1 多模态数据构成

文博语料库宜包含以下类型的多模态数据，以全面、立体地记录和描述文化遗产对象及其背景信息。

5.1.1 文本数据

构成语料库的核心组成部分，包括但不限于：

a) 文物描述文本：如博物馆藏品管理系统中的编目信息、著录条目、鉴定意见、官方网站藏品介绍等；以及不可移动文物的“四有”档案（有保护范围及建设控制地带、有标志说明、有记录档案、有保管机构）、文物保护单位档案、考古工作日记、勘察报告等。
b) 学术研究文献：考古报告、研究专著、学术期刊论文、学位论文等。
c) 历史文献与档案：古籍、地方志、金石拓片题跋、信札、档案文件等。
d) 展览与教育文本：展览前言、单元说明、展品说明牌、遗址现场说明牌、导览词脚本、教育手册、宣传材料等。
e) 口述史与访谈转写文本：对捐赠者、艺术家、修复专家、考古工作者、当地居民、相关社区成员等的访谈记录。

5.1.2 图像数据

直观呈现文物形态、工艺、纹饰、结构、环境的关键信息，包括但不限于：

a) 文物图像：可移动文物的多角度、高分辨率整体与细节照片；不可移动文物的整体全景照片、建筑立面与结构照片、构件细部照片、遗址地貌与探方照片、壁画与石刻细部照片等。
b) 线图与拓片：用于记录纹饰、铭文、建筑平面/立面/剖面图、遗址分布图、探方平面图等的专业图纸。
c) 历史与档案图像：与文物相关的历史照片、手稿扫描件、设计图纸、历史测绘图纸、航空摄影照片等。
d) 情境图像：文物在展览、出土或使用环境中的照片；不可移动文物在其历史与现存环境中的照片。

5.1.3 音视频数据

记录动态过程与声音信息，包括但不限于：

a) 音频：口述史录音、专家讲座、语音导览、遗址环境声、相关音乐或声音景观、播客等。
b) 视频：文物修复过程记录、虚拟展览漫游、专家访谈、考古发掘过程记录、建筑测绘过程、保护工程实施过程、相关纪录片、教育短片等。

5.1.4 三维与交互式数据

记录文物的空间信息与交互体验，包括但不限于：

a) 三维模型：通过激光扫描、结构光扫描或摄影测量技术获取的可移动文物高保真三维数字化模型;以及不可移动文物的建筑、遗址、石窟、大型石刻等的三维数字化模型、数字高程模型(DEM)、点云数据等。
b) 交互式数据：虚拟现实（VR）、增强现实（AR）应用中的交互日志，在线展览、数字遗址公园的用户行为数据等。

5.2 元数据

描述上述各类数据的背景、技术和管理信息的数据，是实现数据发现、管理、理解和互操作的基础。宜包含描述性元数据（如标题、创作者、主题）、技术性元数据（如格式、分辨率、大小）、管理性元数据（如权限、来源）和结构性元数据（如文件组织关系）。

5.3 核心实体与关系

文博语料库的数据组织宜基于实体-关系（E-R）模型或知识图谱思想，明确定义领域核心实体及实体间的语义关系，以构建结构化的文化遗产知识网络。

5.3.1 核心实体类型

核心实体类型宜包括：

a) 物质文化遗产对象包括：

1) 可移动文物：具体的器物、书画、典籍、档案等物理实体。属性包括唯一标识符、名称、分类、材质、尺寸、年代、收藏机构、保存状况等。
2) 不可移动文物：古遗址、古墓葬、古建筑、石窟寺、石刻、近现代史迹等。属性包括唯一标识符、名称、类型、地理位置（坐标）、年代、占地面积、保存状况、保护级别、所属考古学文化等。

b) 行动者：与文化遗产相关的个人或组织，如创作者、收藏者、捐赠者、修复师、建造者、使用者、考古学家、策展机构、文物保护管理机构等。属性包括名称、类型、角色、生卒年份、活动地点等。
c) 时间：描述事件发生或对象存在的时段，如朝代、年代、具体日期、时间段、考古学文化期等。
d) 地点：与文化遗产相关的地理位置，如可移动文物的出土地、收藏地（博物馆）、不可移动文物的所在地址、历史地理位置、发现地、制作地、相关历史地点等，宜采用标准地理编码。
e) 事件：描述与文化遗产相关的活动或过程，如制作、建造、发掘、收藏、展览、修复、勘察、测绘、保护工程、交易等。属性包括事件类型、时间跨度、地点、参与者等。
f) 概念：用于描述和分类的受控术语或类型，如材质（青铜、木、石）、工艺（失蜡法、榫卯、夯筑）、风格（巴洛克、唐代建筑风格）、文物类型（青铜器、书画、古遗址、木构建筑）等。建议链接至权威叙词表。

5.3.2 核心语义关系

核心语义关系宜在数据模型或标注体系中予以定义和实例化，例如：
——（行动者）创作/建造了（物质文化遗产对象）
——（物质文化遗产对象）制作/建造于（时间）
——（可移动文物）出土于（地点）
——（不可移动文物）位于（地点）
——（行动者）捐赠了（可移动文物）给（机构）
——（物质文化遗产对象）参与了（事件）
——（物质文化遗产对象）具有材质（概念）
——（文本/图像资源）描述了（物质文化遗产对象/事件）

通过明确定义实体和关系，并遵循或映射至国际标准本体（如CIDOC CRM, CRMarchaeo扩展），可实现数据的语义化关联和跨系统的互操作。

6 文博语料数据采集

6.1 数据来源

文博语料采集宜系统性地覆盖以下多源、多模态渠道，以确保语料的全面性与立体性。

6.1.1 机构内部权威资源

作为最核心、最权威的数据来源，宜包括：
a) 文博机构的藏品管理系统（CMS）、不可移动文物档案管理系统、数字资产管理系统（DAMS）中的结构化数据与关联文件。
b) 机构内部的研究档案、展览档案、修复记录、考古发掘报告、文物保护工程档案、库房日志及未公开的学术报告。
c) 机构出版的展览图录、学术专著、内部刊物及产生的策展文本、教育材料。

6.1.2 公开出版物与学术数据库

作为高质量、结构化知识的主要来源，宜包括：
a) 学术文献：公开发表的考古报告、研究论文、学位论文。
b) 历史文献与档案：已完成数字化的古籍、地方志、金石拓片、历史档案、报刊等。
c) 权威参考资源：专业百科全书、叙词表、标准文献及政府公开的文化遗产名录、普查报告。

6.1.3 互联网开放资源

作为获取当代语境和公众认知补充，在严格遵守相关法律法规和网站协议的前提下，可采集：

a) 文博机构官方网站发布的新闻、公告、藏品介绍、虚拟展览内容。
b) 专业论坛、博客、社交媒体上关于文博话题的讨论、评论与用户生成内容。
c) 开放存取的学术资源与数据集。

6.1.4 数字化生产与转化

对尚未数字化的物理载体或非文本资源进行转化，包括：

a) 文本数字化：对纸质文献、档案进行高精度扫描，并利用OCR技术转换为可处理文本，辅以人工校对。
b) 音视频转写：对专家讲座、口述史访谈、修复过程录像等音频、视频资料进行人工或ASR辅助转写，生成文本副本。
c) 三维数据采集：通过激光扫描、摄影测量等技术，对文物、遗址、建筑进行三维数字化，生成点云、网格模型及纹理信息。

6.2 采集原则与策略

文博语料采集宜遵循以下原则，并制定相应策略，以确保语料库的基础质量、合法性与可持续性。

6.2.1 系统性与代表性

采集工作宜有计划、成体系地进行，确保语料在时间跨度（如不同历史时期）、地域分布、文物门类（如陶瓷、书画、青铜器）、信息类型（学术、教育、管理）和数据模态（文本、图像、音视频、3D模型）上具有足够的覆盖度和代表性，避免样本偏差。

6.2.2 质量优先与权威性

建立数据源准入评估机制，优先采集来源可靠、内容准确、学术价值高的语料。对于学术文献，宜优先选择核心期刊、权威出版机构；对于机构数据，宜以官方发布为准。

6.2.3 合法合规与伦理安全

采集前必须进行版权状态审查与授权获取。对于受版权保护的材料，需获得权利人明确授权；对于互联网公开内容，应遵守相关协议要求，设置合理抓取频率，避免对目标服务器造成负担。

涉及个人隐私（如口述史）的数据，须进行匿名化处理，并获取知情同意。整个采集过程应符合《数据安全法》、GDPR等数据安全与隐私保护相关法律法规。

6.2.4 原始性保留与可追溯性

在采集和初始存储阶段，宜尽可能保留数据的原始格式、元数据及上下文信息（如网页原始HTML、文献的版式信息）。所有语料均宜记录明确的来源标识、采集时间和采集方式，确保数据的可追溯性与审计能力。

6.2.5 平衡性与动态性

在保证核心、权威语料的基础上，注意平衡不同观点、不同来源的语料。建立持续采集与更新机制，对新闻、学术动态等时效性强的语料进行定期增量采集，使语料库能够反映知识的发展与更新。

7 文博语料数据清洗

7.1 清洗流程与要求

对采集的原始语料（“生语料”）宜进行系统化、标准化的清洗与预处理，以去除噪声、纠正错误、统一格式，形成高质量、结构化的“熟语料”。主要流程与要求如下：

7.1.1 格式统一与转换

将来自不同源头、格式各异的异构数据（如DOC、PDF、HTML、数据库记录）统一转换为便于后续处理的内部标准格式，如纯文本（TXT）、结构化XML或JSON格式。

7.1.2 内容去噪

使用正则表达式、基于规则或机器学习的方法，去除文本中的HTML标签、广告、导航栏、版权声明、页眉页脚等与核心内容无关的“样板”信息。

7.1.3 去重

在不同粒度上识别并移除重复或高度相似的内容，确保语料库的唯一性。包括：

a) 文件级去重：通过计算文件哈希值（如MD5、SHA-1）识别并删除完全相同的文件。
b) 段落/句子级去重：将文档拆分为段落或句子单元，识别并移除内容重复的单元。
c) 近似去重：使用SimHash、MinHash等局部敏感哈希（LSH）算法，高效识别并处理内容高度相似但非完全相同的文本。

7.1.4 错误纠正

识别并修正数据中的各类错误。包括：

a) 技术性错误：自动或半自动地纠正OCR识别产生的字符错误、乱码。
b) 规范性错误：修正明显的拼写错误、标点误用。
c) 知识性错误：对于可通过权威知识库校验的明显事实错误（如错误的朝代纪年），应予以标注或修正。

7.1.5 文本规范化

将形式不同但意义相同的文本单元转换为统一的标准表示形式。包括：

a) 字符处理：进行繁简中文转换、全角/半角字符转换、特殊符号（如引号、破折号）标准化。
b) 编码归一化：将所有文本统一转换为UTF-8编码，避免乱码。
c) 格式规范化：处理多余的空格、制表符、换行符，将连续的空白符合并为单个空格。

7.1.6 数据脱敏

对语料中可能包含的个人隐私信息（如口述史中的姓名、住址）、敏感位置信息（如未公开的考古遗址精确坐标）、内部操作信息等进行脱敏处理。处理方法包括但不限于替换、泛化、屏蔽或加密。

7.1.7 数据整合与关联

将来自不同数据源、描述同一文物、事件或概念的文本、图像等信息进行初步关联和对齐，为后续构建知识关联奠定基础。

7.2 质量控制

宜建立贯穿清洗与预处理全过程的质量控制机制，确保数据的纯净度、准确性与一致性。

7.2.1 流程化与自动化

宜建立自动化的数据处理管道，使用工作流调度工具串联各清洗步骤，确保所有数据经过统一、可复现的处理流程。

7.2.2 多级校验

对数据进行多级校验，包括：

1) 自动化脚本校验：在每个关键清洗步骤后，部署自动化脚本对处理结果进行初步验证，例如检查格式转换成功率、去重比例、字符编码一致性等。

2) 人工抽样审核：对自动化处理的结果，尤其是错误纠正、实体关联等复杂环节，必须进行定期的人工抽样审核。审核宜由具备文博领域知识的人员执行，以确保处理结果的准确性。

7.2.3 质量记录

记录每次清洗作业的关键参数、处理结果统计（如去除了多少噪声、修正了多少错误）以及发现的主要问题，用于流程优化和版本追溯。

8 文博语料数据标注

8.1 标注体系设计原则

文博语料库的标注体系是其从原始数据转变为结构化、可计算知识资源的核心环节。设计该体系宜遵循以下原则，以确保标注工作的科学性、规范性、可扩展性与实用性：

a) 层次性与系统性：标注体系宜设计为多层次结构，从宏观的篇章功能到微观的词法句法，从通用的语言学信息到深度的领域知识，逐层递进，系统覆盖。
b) 标准化与一致性：宜优先采用或兼容成熟的国际、国内或行业标注规范。必须制定详尽、无歧义的《标注指南》，并定期进行标注者间一致性评估，确保不同人员、不同批次的标注结果具有高度一致性。
c) 领域相关性：标注体系必须紧密结合文博领域的知识特性，设计专用的实体类型（如文物、工艺、朝代）、关系类型（如“创作于”“出土于”）和事件框架，以准确捕捉领域语义。
d) 可扩展性：体系架构宜具备良好的开放性，能够根据新的研究需求、数据类型或技术发展，方便地增加新的标注层级、标签类型或属性，而无需重构整体框架。
e) 工具友好性与人机协同：标注格式宜便于主流标注工具和自然语言处理框架解析与处理。鼓励采用“AI预标注+人工校验与修正”的人机协同模式，以提高大规模标注的效率与质量。
f) 多模态关联性：标注体系需支持并明确建立文本、图像、音频、视频、三维模型等不同模态数据之间的语义关联与对齐，为实现跨模态检索与理解奠定基础。

8.2 多层次标注内容

文博语料库的标注宜采用分层、多维的架构，具体内容如下：

8.2.1 基础语言学标注层

此层为所有文本语料提供通用的语言学信息标注，是上层领域分析的基础。

a）分词与词性标注：对中文文本进行词语切分，并为每个词语标注其词性（如名词、动词、形容词等）。

b）命名实体识别：识别并标注文本中的通用命名实体，主要包括人名、地名、组织机构名、时间表达式等。

c）句法分析：分析句子的语法结构，可进行成分句法分析或依存句法分析，揭示词语间的句法关系。

d）语义角色标注：标注句子中谓词（通常是动词）与其相关论元（如施事、受事、时间、地点等）之间的语义关系。

8.2.2 领域知识标注层

8.2.2.1 核心文博实体识别与分类

在命名实体识别基础上，扩展识别文博领域的特有实体类型，建议包括但不限于：

——具体的物质文化遗产对象（如“可移动文物：后母戊鼎”“《清明上河图》”；不可移动文物：“长城”“秦始皇陵”“应县木塔”“莫高窟第 45 窟”）。
——文物材质（如“青铜”“玉石”“砖木”）。
——制作工艺或技术（如“失蜡法”“釉下彩”“抬梁式构架”“壁画地仗层制作工艺”）。
——朝代、历史时期或文化期（如“唐代”“新石器时代仰韶文化”）。
——历史人物（创作者、收藏者、建造者、相关人物）。
——地理位置（可移动文物的遗址、出土地、不可移动文物的所在地、历史沿革地点、收藏机构）。
——历史或文博相关事件（如“制作”“建造”“发掘”“展览”“修复”“保护工程”）。
——艺术风格、流派或类型（如“青花瓷”“文人画”“唐代石窟艺术”“闽南民居风格”）。
——考古学文化（如“龙山文化”“二里头文化”）。
——建筑构件与结构（如“斗拱”“柱础”“阙”“封土”）。

8.2.2.2 关系与事件抽取

在核心文博实体识别与分类的基础上，进行关系和时间抽取，包括：

a) 识别并标注上述实体之间的语义关系，形成“（实体A，关系，实体B）”的三元组。
b) 识别文本中描述的复杂事件，并标注其类型（如制作、发掘）及事件要素，包括参与者、时间、地点、工具等，遵循以事件为中心的知识组织思想。

8.2.2.3 篇章与功能标注

对篇章和功能进行标注，包括：

a) 文本类型标注：标识语料的体裁或来源，如“考古报告”“展览说明”“学术论文”“新闻稿”“导览词”。
b) 段落/句子功能标注：标注文本单元在篇章中的功能，如“形制描述”“工艺分析”“历史背景介绍”“价值评述”“作者观点”。
c) 情感与立场标注：对于评论文、展览叙事等文本，可标注作者或叙述者对某一对象（文物、人物、事件）的情感倾向（正面、负面、中性）或特定立场。

8.2.3 多模态关联标注

为实现多模态数据的融合与互操作，需进行跨模态关联标注。

8.2.3.1 跨模态对齐与关联

跨模态对齐与关联包括：

a) 图文对齐：建立文本描述片段与图像中特定区域（通过边界框、多边形分割标注）的对应关系。例如，将描述“瓶身绘有缠枝莲纹”的文本与图像中瓶身纹饰区域关联。
b) 音视频-文本对齐：将访谈、讲座等音频/视频的转写文本与媒体时间轴进行精确对齐。
c) 文本-3D模型关联：将描述文物部件、结构或修复部位的文本与三维模型上的特定网格或区域进行关联。

8.2.3.2 非文本模态的独立标注

非文本模态的独立标注包括：

a) 图像标注：对图像进行对象检测（框出文物）、实例分割（精确分割文物轮廓）、关键点检测（标注特定结构点）或打上描述性标签。
b) 3D模型标注：在三维模型上标注部件、损伤部位、铭文位置、测量点等。

8.3 标注流程、规范与工具

8.3.1 标注流程管理

标注工作宜遵循系统化的流程，通常包括：需求分析与《标注指南》制定、标注员培训与考核、试标注与指南迭代、正式标注与过程质检、多轮审核（标注员自查、交叉校验、专家仲裁）、数据清洗整合、最终交付与版本管理。

8.3.2 标注规范制定

宜编制详细的《文博语料标注指南》，明确所有标签的定义、标注规则、边界案例处理、正反示例，并随项目进展迭代更新。

8.3.3 标注工具选型

宜选用功能成熟、支持协作与项目管理的标注平台或工具，后台宜支持：任务分配、进度跟踪、多人协同、版本管理、结果导出等功能。

8.3.4 质量控制与评估：

对质量进行控制和评估，包括但不限于以下方法：

a) 标注者间一致性评估：定期计算不同标注员对同一批语料的标注一致性。对于一致性低的项目，需复盘原因并重新培训或修订指南。

b) 人工审核与仲裁：建立多级审核机制，最终由领域专家对存疑或困难的标注案例进行仲裁，确保标注结果的权威性与准确性。
c) 自动化辅助校验：开发或利用脚本对标注结果的格式合规性、标签使用规范性进行自动检查。

8.4 标准化映射与语义发布

为提升语料库的互操作性与长期价值，鼓励将标注成果映射至国际通用的语义标准，包括：
a) 元数据映射：将语料及其标注结果的描述性元数据，映射至文博领域广泛应用的交换标准。
b) 知识模型映射：将标注出的核心文博实体、属性及关系，按照定义的映射规则，转换为以CIDOC CRM本体为基础的知识模型。
c) 语义化发布：将基于本体的知识模型，以RDF三元组的形式进行编码和存储，并为每个实体分配可解析的URI。最终，可将部分或全部数据以链接开放数据的形式发布，实现与全球文化遗产数据网络的互联，支持复杂的语义查询与推理。

9 文博语料测试

9.1 通用测试内容

9.1.1 输入输出规范性检测

验证语料存储格式是否符合领域通用标准，检查核心元数据的完整性与格式规范性。

9.1.2 文博概念准确性检测

通过文博领域知识图谱、权威叙词表或本体库，匹配语料中的专业术语，验证其使用的准确性及上下文一致性。

9.1.3 多模态关联性检测

测试跨模态数据的对齐能力，确保图文描述对应、音视频转写文本与时间戳同步、三维模型部件与文本描述的关联准确有效，支持跨模态检索与应用的协同。

9.1.4 模型输入压力检测

将语料输入目标文博大模型或应用系统，监测其处理过程中的异常输出、逻辑矛盾或事实性错误，并溯源至语料自身的缺陷，识别隐含的噪声、标注错误或知识冲突。

9.1.5 语料分布及覆盖性检测

统计分析语料在时间线（朝代）、地域、文物门类、文物等级、数据模态等维度的覆盖度与分布比例，确保其符合实际文化遗产资源分布与研究需求，并评估关键但罕见的样本是否已被纳入。

9.1.6 语料价值对齐检测

测试文博语料库所隐含的文化价值观与历史观，确保语料内容符合科学伦理、历史文化事实、社会道德及正确的文化传承导向，避免传播历史虚无主义或文化偏见。

9.2 学科特色测试

9.2.1 文博领域知识图谱验证

文博领域知识图谱验证包括：
a) 检查基于语料构建或关联的知识图谱的拓扑结构合理性，包括文博实体（人、物、地、时、事）关系路径的连通性、属性值完整性、层级关系逻辑性。
b) 验证实体关系（如“创作于”“出土于”“收藏于”）是否符合领域共识与历史事实，属性值（如年代、尺寸）范围合理。
c) 对来自不同数据源（如不同博物馆、文献）的冲突信息进行实体对齐与关系消歧，确保知识的一致性。

9.2.2 文博知识动态更新验证

文博语料库宜建立机制，跟踪考古新发现、学术研究进展及知识版本的迭代，测试语料库在纳入新知识时的逻辑延续性，并记录完整的变更溯源链。

9.2.3 行业应用有效性检测

定期选择典型文博机构、研究团队或应用开发商作为试点，调查语料库在其实际业务场景中的可用性、易用性与有效性，收集反馈并用于优化语料库。

9.3 人工检测

9.3.1 人工检测流程

文博语料的人工检测主要包括以下内容：

a) 由具备基础文博知识的初级检测员执行初审，检查语料的格式错误、明显的事实矛盾，并进行一致性抽查，生成问题清单。
b) 同一批语料由至少两位具备丰富文博经验的高级检测员独立进行交叉审核，重点核查语料的逻辑性、专业事实准确性及复杂语义标注的质量。若检测结果一致性低于既定阈值，则启动专家仲裁流程。
c) 由文博领域专家或资深研究员对交叉审核中的分歧样本进行最终仲裁，并对高频错误类型进行分析，提出针对性的语料优化与标注指南修订建议。

9.3.2 检测人员资质要求

文博语料检测需结合领域知识和技术能力，对检测人员进行分层管理：

a) 初级审核员：须具备文博、历史、考古等相关学科基础教育背景，或1至3年文博领域从业/研究经验。
b) 高级审核员：须具备3年以上文博领域专业工作经验，持有相关专业能力认证，可主导复杂语料的交叉审核。
c) 专家审核员：需在文博特定领域具备专业权威性，拥有高级职称或博士学历，且具备5年以上行业内深入研究或管理经验，并发表过相关领域学术成果或参与过标准制定，负责终审仲裁与技术指导。

10 文博语料使用

10.1 学术研究与文化遗产保护

10.1.1 智慧科研辅助

基于高质量、语义关联的文博语料库，为历史、考古、艺术史、建筑史、文物保护科学等学科研究者提供数据驱动的分析基础。支持基于知识图谱的复杂关系查询（如器物演变序列、人物社会网络、文物传世路径）、文献计量与知识挖掘，从而发现潜在知识关联，提出新的研究假设，革新人文社科学术研究方法。

10.1.2 文物保护专业辅助

以高精度文物图像（含多光谱/高光谱数据）、三维模型、历史修复档案、考古报告及材料分析数据为核心语料，构建专业知识库及分析模型，赋能可移动与不可移动文物数字化存档、病害分析、修复方案制定、风险建模与虚拟仿真等全流程，兼顾文物保护决策的专业严谨性与公共科普的可及性，有效提升文物预防性保护、修复及修缮工作的效率与科学性。

10.1.3 行业数据基础设施与行政决策辅助

作为可共享、可互操作的高质量数据资产，构建国家级或区域级涵盖可移动与不可移动文物的文博知识网络。相关数据在科研辅助外，可为文物普查、保护规划编制、文物保护工程管理、安全监测等业务提供数据支撑和知识服务。

10.2 公共服务与文化体验

10.2.1 智能导览与知识问答

融合可移动与不可移动文物多模态描述数据、相关学术文献、策展大纲、专家讲解词、公众问答对等高质量语料，构建可计算的文博“知识大脑”，驱动智能导览系统。相关系统可服务于博物馆、考古遗址公园等场景，为游客提供7×24小时、个性化、多轮对话的自然语言问答与讲解服务，解决讲解资源时空分布不均、知识深度与广度不足的痛点，变单向信息输出为双向交互探索，实现“千人千面”的个性化体验。

10.2.2 在地文化创新体验

聚焦年轻群体对在地文化与历史知识趣味性探索的需求，将文博语料（如城市历史沿革、名人轨迹、老照片、口述史）转化为CityWalk（城市漫步）剧本游的叙事线索、解密任务与互动环节，将城市漫步与实景解谜、知识问答相结合，打造个性化、可复用的文化深度体验产品。

10.2.3 沉浸式线下体验构建

聚焦VR大空间、主题体验馆、沉浸式特展等线下体验场景，将文博语料（如古建点云数据、文物数据、历史情境描述）转化为可感知、可交互的空间化与叙事化内容，为虚拟场景渲染、物理动线规划、互动环节设计提供精准的文化逻辑与细节依据，使用户在自由探索与互动参与中获得“置身历史现场”的深度沉浸感，实现从“被动观看”到“主动探索”的体验升级。

10.3 数字内容高质量生成

10.3.1 多模态模型训练

将体系化、标注清晰的文博图像语料（如器物、书画、纹饰、古建）用于训练图像、视频生成类AI大模型。应用宜聚焦中华优秀传统文化核心视觉符号（如传统色彩、纹样、构图法则），通过语料注入解决通用模型生成“中国元素”内容时常见的风格失真、细节粗糙、历史语境错位等问题，将静态馆藏资源转化为AI模型的“文化基因”，支撑符合历史语境和中式审美意趣的数字化内容规模化生产。

10.3.2 影视与数字内容考证设计

将文博语料系统性植入历史题材影视剧、动画、纪录片的前期制作流程，为服装、化妆、道具（服化道）及场景陈设的设计提供精准的考据依据与创意来源，赋能美术指导、造型师等从业人员，显著提升作品的历史还原度、视觉真实感与文化内涵，有效避免“穿越”式错误，增强作品的历史厚重感与中式审美意趣。

10.3.3 游戏及 XR 内容开发

将文博语料深度融入主题游戏、动画及扩展现实（VR/AR/MR）体验项目的世界观设定、虚拟场景建模、角色与道具设计、玩法机制及非玩家角色（NPC）对话逻辑设计中。应用宜聚焦提升数字内容的文化真实性与用户沉浸感，从源头规避朝代特征错配、艺术风格割裂等问题，让用户、玩家在探索与互动中自然习得历史文博知识，实现寓教于乐的文化主动传播。

11 数据安全、伦理与治理

11.1 版权与数据许可

11.1.1 版权状态审查与授权

在数据采集前，应系统审查语料的版权状态。对于明确受版权保护的材料（如当代出版物、未公开档案），应依法取得权利人的明确授权。对于已进入公共领域的材料（如古籍），应予以明确标识。在学术研究等特定场景下，应审慎评估并遵循“合理使用”原则，但需注意其法律风险。

11.1.2 数据许可策略制定

应制定清晰、分层的数据许可与访问策略。鼓励采用国际通行的知识共享（Creative Commons）系列协议发布可公开共享的数据，并根据数据敏感性与价值，设计开放访问、注册访问、受控访问等多级权限。所有许可条款或使用协议应在平台显著位置明确公示。

11.1.3 语料库知识产权声明

经过创造性选择、编排、标注而形成的语料库整体，可作为汇编作品享有相应版权。建设方应明确声明其对语料库（作为数据库）的整体权利。

11.2 隐私与伦理保护

11.2.1 个人隐私信息处理

在采集和处理涉及个人身份信息的数据（如口述史、访谈录、信札）时，必须严格遵守《个人信息保护法》等相关法规。必须对姓名、住址、联系方式等敏感信息进行彻底的匿名化或假名化处理，确保无法识别特定个人。

11.2.2 文化敏感性与社群权利尊重

对于涉及特定民族、宗教社群或原住民的文化遗产数据，其采集、标注、使用与传播应充分尊重相关社群的文化权利与意愿。宜建立伦理审查机制，或在涉及重大文化敏感性内容时，与相关社群代表进行协商。

11.3 数据安全与存储

11.3.1 全生命周期安全防护

应建立覆盖数据采集、传输、存储、处理、销毁全生命周期的安全防护体系。技术措施包括但不限于：数据传输使用HTTPS等加密通道；存储数据采用加密技术；部署防火墙、入侵检测等安全防范与监控系统；实施严格的访问控制与身份认证机制。

11.3.2 数据备份与灾难恢复

宜建立可靠的数据备份与灾难恢复机制，实施定期备份和异地容灾，确保数据的持久可用性。应定期进行恢复演练，验证备份的有效性。

11.4 可持续治理与生态发展

11.4.1 持续维护与更新机制

语料库宜建立持续的维护制度，包括定期采集新数据、根据反馈修正错误、优化标注体系。必须实施严格的版本控制（如使用Git、DVC等工具），记录所有重大变更，保障研究的可复现性。

11.4.2 组织与财务可持续性

宜由建设方、合作机构、领域专家、技术专家等组成的治理委员会，负责制定发展战略、协调资源、监督运营。宜探索多元化的可持续资金支持模式，如项目经费、机构预算、合作开发、公益资助等。

11.4.3 社区参与与生态构建

宜通过众包模式吸引专业社区参与数据校对、标注等工作。宜提供完善的开放API接口和开发文档，支持基于语料库的二次开发和创新应用。通过建立用户社区、举办开发者活动等方式，构建活跃的语料库应用生态。

参 考 文 献

[1] DB11/T 1219—2015 文物艺术品元数据规范
[2] WH/T 66—2014 古籍元数据规范
[3] WW/T 0114—2023 可移动文物二维数字化采集与加工
[4] WW/T 0115—2023 可移动文物三维数字化采集与加工
[5] 刁常宇. 有器之用：馆藏文物数字化采集与质量评价[M]. 杭州: 浙江大学出版社, 2021.
[6] 庄颖. 面向人工智能的博物馆藏品知识组织——以故宫博物院“中国古代可移动文物概念参考模型”为例[J]. 故宫博物院院刊,2023,(11):126-136+150.DOI:10.16319/j.cnki.0452-7402.2023.11.004.
[7] 叶祎珮. “中国古代可移动文物概念参考模型”构建实践[J].数字人文研究,2023,3(03):37-48.
[8] 李琳. 基于知识图谱的文物数字化系统构建研究[D]. 北京交通大学,2022.DOI:10.26944/d.cnki.gbfju.2022.002988.
[9] 张敏. 面向文物领域的知识图谱构建技术研究[D]. 西北大学,2021.DOI:10.27405/d.cnki.gxbdu.2021.000022.
[10]张娜.文物知识图谱构建关键技术研究与应用[D].浙江大学,2019.
[11]杨伟强.文物知识图谱的构建与应用[D].天津大学,2018.
[12]林炀平.文物知识图谱构建与检索关键技术研究与实现[D].浙江大学,2017.